การเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกข้อมูลโรคหัวใจด้วยเทคนิคเหมืองข้อมูล

คมสัน แสนรักษ์[1] ,ศิริพรรณ กันทะมูล[1] และ ไก้รุ่ง เฮงพระพรหม[1*]

[1] *สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม*

[*] *kairung2011.heng@gmail.com*

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทขอมูล 2) ศึกษาผลการทดลองใช้โปรแกรม RapidMiner Studio ประกอบดวยเทคนิคการจำแนกขอมูล 3 เทคนิค ไดแก่ วิธีเพื่อนบ้านใกล้ที่สุด, เบย์อย่างง่าย, เครือข่ายประสาท โดยทำการทดสอบกับขอมูลโรคหัวใจจำนวน 303 คน

ผลการวิจัยพบว่า การเปรียบเทียบประสิทธิภาพการจำแนกประเภทขอมูลเฉลี่ยดีที่สุดที่ คือ Neural Network มีความแม่นยำมากที่สุดคือ 82.17% รองลงมาคือ Naïve Bayesian 81.88% และ K-Nearest Neighbor 65.66% โดยใชจำนวนคุณลักษณะเพียง 11 คุณลักษณะ โดยที่เทนิค Nerve net มีความแม่นยำมากที่สุด และการเปรียบเทียบประสิทธิภาพแสดงใหเห็นวาการคัดเลือกคุณลักษณะที่เหมาะสมกอนการจำแนกประเภทขอมูลนั้นเปนแนวทางที่เหมาะสมอยางยิ่งที่จะทำใหเพิ่มประสิทธิภาพของการจำแนกขอมูลไดเปนอยางดี จึงจำเปนอยางยิ่งที่จะตองทำการกรองขอมูลกอนเบื้องต้น

คำสำคัญ: *เหมืองข้อมูล, วิธีเพื่อนบ้านใกล้ที่สุด, เบย์อย่างง่าย,เครือข่ายประสาท*

A Comparison of the Efficiency of Heart Disease Data Classification using Data Mining Technique

Komsan Sanrak [1], Siripun Kuntamoon [1] and Kairung Hengpraphrom [1*]

[1] *Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University*

[*] *kairung2011.heng@gmail.com*

ABSTRACT

The purposes of the research were to 1) to compare the technical features, select the most suitable properties 2) study the experimental results using RapidMiner Studio program, consisting of condensed calculation techniques, 3 most techniques, simple ben, neural net by testing with the heart 303 people

The research findings showed that the comparison of the efficiency of the average data classification was Nerve net with the most accuracy of 82.17%, followed by Naïve Bayesian 81.88% and K-Nearest Neighbor 65.66%. By using only 11 features. With Nerve net technology being the most accurate and the performance comparison shows that the selection of appropriate characteristics before data classification is an appropriate way to enhance the efficiency of data classification. Therefore, it is very necessary to first filter the data.

Keyword: Data mining, K-Nearest Neighbor, Naïve Bayesian, Neuron network

1. บทนำ

โรคหัวใจ เป็นสาเหตุอันดับที่ 2 ของการเสียชีวิตในประเทศไทย โรคหัวใจ (Heart Disease) หมายถึง โรคต่าง ๆ ที่ส่งผลกระทบต่อการทำงานของหัวใจ โดยโรคหัวใจสามารถแบ่งย่อยได้เป็นหลายกลุ่มโรค เช่น โรคหลอดเลือดหัวใจ โรคกล้ามเนื้อหัวใจ โรคหัวใจเต้นผิดจังหวะ โรคลิ้นหัวใจ โรคหัวใจพิการแต่กำเนิด และโรคติดเชื้อบริเวณหัวใจ เป็นต้น นับเป็นปัญหาทางสาธารณสุขที่สำคัญของทั้งประเทศไทยและทั่วโลก ก่อให้เกิดผลเสียต่อสุขภาพร่างกายประชาชน จากรายงานขององค์การอนามัยโลก (WHO) ในปี 2559 พบทั่วโลกมีผู้เสียชีวิตปีละประมาณ 57 ล้านคน โดยโรคหัวใจขาดเลือดเป็นสาเหตุการเสียชีวิตอันดับ 1 มีผู้เสียชีวิตประมาณ 9.2 ล้านคน สำหรับประเทศไทยในปี 2560 พบว่าคนไทยมีแนวโน้มการป่วยด้วยโรคหัวใจขาดเลือดถึง 326,946 คน และเสียชีวิตจากโรคหัวใจขาดเลือดถึง 20,746 คน เฉลี่ยเสียชีวิตวันละ 57 คน และมีแนวโน้มการป่วยเพิ่มขึ้นอย่างต่อเนื่อง

สาเหตุการเกิดโรคหัวใจ โรคหลอดเลือดหัวใจ มักมีอาการเจ็บหรือแน่นหน้าอก ร้าวไปตามกราม แขน ลำคอ ท้อง หรือบริเวณหลัง และอาจจะมีอาการเหนื่อย อ่อนเพลีย ไม่มีแรง หรือหมดสติได้ สาเหตุเกิดจากไขมันหรือแคลเซียม ที่สะสมมากขึ้นเรื่อย ๆ ในหลอดเลือด จนขัดขวางทางเดินเลือด ทำให้เลือดไปเลี้ยงกล้ามเนื้อหัวใจ ไม่เพียงพอ และอาจจะมีสาเหตุอื่น ๆ เช่น โรคเบาหวาน ความดันโลหิตสูง ไขมันในหลอดเลือดสูง น้ำหนักเกิน และสูบบุหรี่, โรคหัวใจเต้นผิดจังหวะ หัวใจอาจจะเต้นเร็วผิดปกติ ช้าผิดปกติ หรือเต้นไม่สม่ำเสมอ ทำให้รู้สึกเหมือนใจสั่น มันแสดงอาการเหนื่อยง่าย แน่นหน้าอก เวียนศีรษะ หรือคล้ายจะเป็นลมได้ สาเหตุอาจจะเกิดจากการถูกไฟฟ้าช็อต การใช้สารเสพติด ยาอาหารเสริมบางชนิด รวมทั้ง ดื่มแอลกอฮอล์และคาเฟอีน นอกจากนี้ อาจจะมีความเสี่ยงจากอาการเจ็บป่วย โรคหัวใจอื่น ๆ เช่น โรคหัวใจพิการแต่กำเนิด โรคหลอดเลือดหัวใจ โรคลิ้นหัวใจ หรือโรคอื่น ๆ เช่น เบาหวาน ความดันโลหิตสูง, โรคกล้ามเนื้อหัวใจ มักมีอาการเหนื่อยง่าย หายใจไม่อิ่ม เมื่อต้องออกแรงมาก ๆ มักมีอาการมากขึ้น หากมีอาการรุนแรงมากแม้ขณะนั่งอยู่เฉยๆ ก็จะมีอาการเหนื่อย บวมตามแขน ขา หนังตา ร่วมกับอาการเพลีย นอนราบไม่ได้ และไอตื่นขึ้นมาไอในเวลากลางคืน สาเหตุของการเกิดโรคกล้ามเนื้อหัวใจ สาเหตุมักจะต่างกันไปตามความผิดปกติที่เกิดขึ้นกับกล้ามเนื้อหัวใจ โรคกล้ามเนื้อหัวใจเสื่อม ทำให้การไหลเวียน ของเลือดสู่หัวใจน้อยลง การได้รับยาบางชนิด การติดเชื้อ และพันธุ์กรรม และอาจจะเป็นผลจากโรคอื่น ๆ เช่น เนื้อเยื่อเกี่ยวพันผิดปกติ ภาวะธาตุเหล็กมากเกิน หรือการรักษามะเร็งด้วยเคมีบำบัด รังสีบำบัด เป็นต้น

จากความสำคัญและปัญหาที่กล่าวมาข้างต้น คณะผู้จัดทำจึงเล็งเห็นถึงความสำคัญของโรคหัวใจ และเพื่อจำแนก ประเภทข้อมูลว่ากลุ่มใดบ้างที่เสี่ยงต่อการเป็นโรคหัวใจ และเพื่อทดลองใช้โปรแกรม RapidMiner Studio เพื่อทดสอบประสิทธิภาพในการเปรียบเทียบการจำแนกข้อมูลโดย เทคนิค 3 แบบ คือ K-Nearest Neighbor, Naïve Bayesian, Neural Network

2. วัตถุประสงค์การวิจัย

เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทขอมูล

3. เอกสารและงานวิจัยที่เกี่ยวข้อง

ไพศาล จันทรเจริญ , สุพจน์ เฮงพระพรหม และ ไกรุง เฮงพระพรหม (2559) ได้ทำการวิจัยเรื่อง การเปรียบเทียบประสิทธิภาพการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทขอมูลไมโครอารเรย วิธีการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทขอมูลไมโครอารเรย โดยในการทดลองจะทำการทดสอบ 2 ขั้นตอน คือ 1) การคัดเลือกคุณลักษณะที่เหมาะสม ดวยวิธีCosine และ SNR โดยทำการคัดเลือกคุณลักษณะที่เหมาะสมแบงเปนกลุมยอยๆ ตั้งแต 100 ถึง 1000 คุณลักษณะ และ2) การทดสอบประสิทธิภาพการจำแนกขอมูล ประกอบดวยเทคนิคการจำแนกขอมูล 3 เทคนิค ไดแกตนไมตัดสินใจ วิธีเพื่อนบานใกลที่สุดเค และ เบยอยางงาย โดยทำการทดสอบกับขอมูลโรคมะเร็งตอมน้ำเหลือง จำนวน 47 ตัวอย่าง 4,026คุณลักษณะ ผลของการทดลองพบวา วิธีการคัดเลือกคุณลักษณะแบบ SNR สามารถให ประสิทธิภาพการจำแนกประเภทขอมูลเฉลี่ยดีที่สุดที่94.33% โดยใชจำนวนคุณลักษณะเพียง 100 คุณลักษณะ ขณะที่การจำแนกขอมูลดวยคุณลักษณะทั้งหมดใหประสิทธิภาพที่ 86.52% จึงแสดงใหเห็นวาการคัดเลือกคุณลักษณะที่เหมาะสมกอนการจำแนกประเภทขอมูลนั้นเป็นแนวทาง

สุรวัชร ศรีเปารยะ และสายชล สินสมบูรณ์ทอง (2560) ได้ทำการวิจัยเรื่อง การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง :กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุด วิธีต้นไม้ตัดสินใจ วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติกและวิธีนาอีฟเบย์เพื่อวัดประสิทธิภาพการจำแนกกลุ่ม โดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโลประเทศอินเดีย โดยแบ่งข้อมูลเป็นชุดสร้างตัวแบบ และชุดทดสอบตัวแบบ ในอัตราส่วน 70 และ 30 ตามลำดับ จากการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรัง โดยเปรียบเทียบจากค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยวิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุดคือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้อง คือ 100 % และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

ณัฐวดี หงษบุญมี และ พงศนรินทร ศรรุง (2561) ได้ทำการวิจัยเรื่อง การประยุกตใชเทคนิคจำแนกขอมูลแบบตนไมตัดสินใจเพื่อการวินิจฉัยโรคในโคเบื้องตนบนโทรศัพทมือถือ ถือ โดยทำการรวบรวมขอมูลปจจัยที่เกี่ยวของกับการวินิจฉัยโรคโนโคจากกลุมเกษตรกรผูเลี้ยงโคและผูเชี่ยวชาญในเขตจังหวัดพิษณุโลก สรางโมเดลวินิจฉัยโรคใชเทคนิคตนไมตัดสินใจ เปรียบเทียบอัลกอริทึมจำนวน 3 อัลกอริทึม ไดแก J48 RandomTree และ REPTree แลวทำการทดสอบประสิทธิภาพโมเดลดวยวิธีการตรวจสอบแบบไขว เพื่อที่จะหาโมเดลการวินิจฉัยโรคที่มีประสิทธิภาพดีที่สุดจากผลการวิจัยพบวาตนไมตัดสินใจแบบอัลกอริทึม RandomTree มีประสิทธิภาพดีที่สุด คาความถูกตองเทากับ99.47% คาความคลาดเคลื่อนระหวางคาจริงและค่าที่พยากรณไดเทากับ 0.020 คาความความแมนยำเทากับ 0.995คาความระลึกเทากับ 0.995 และคาความถ

วงดุลเทากับ 0.995 จากนั้นนำอัลกอริทึมที่ให้คาความแมนยำสูงสุดไปพัฒนาเปนระบบวินิจฉัยโรคในโคเบื้องตนในรูปแบบแอพพลิเคชั่นบนโทรศัพทมือถือ ซึ่งสามารถวินิจฉัยโรค แสดงขอมูลรายละเอียดโรค สาเหตุ อาการและการปองกันโรคในโคได ผลการประเมินความพึงพอใจแอพพลิเคชั่นจากผูใชซึ่งเปนกลุมเกษตรผูเลี้ยงโคและผูใชงานทั่วไปจำนวนทั้งหมด 35 คน พบวาผูใชมีความพึงพอใจตอระบบคาเฉลี่ยรวมเทากับ 4.01 คาสวนเบี่ยงเบนมาตรฐานเทากับ 0.55 ดังนั้นจึงสามารถสรุปไดวาแอพพลิเคชั่นที่พัฒนาขึ้นนั้นมีประสิทธิภาพอยูในระดับดี

4. วิธีดำเนินการวิจัย

4.1 ขั้นตอนการดำเนินการวิจัยประกอบด้วย 3 ขั้นตอน ดังนี้

โรคหัวใจ (Heart Disease) หมายถึง โรคต่าง ๆ ที่ส่งผลกระทบต่อการทำงานของหัวใจ โดยโรคหัวใจสามารถแบ่งย่อยได้เป็นหลายกลุ่มโรค เช่น โรคหลอดเลือดหัวใจ โรคกล้ามเนื้อหัวใจ โรคหัวใจเต้นผิดจังหวะ โรคลิ้นหัวใจ โรคหัวใจพิการแต่กำเนิด และโรคติดเชื้อบริเวณหัวใจ เป็นต้น

1.1 ศึกษาและวิเคราะห์ปัญหาที่เกี่ยวข้องกับงานวิจัย

1.2 วิเคราะห์ข้อมูลที่ได้จากการศึกษาในขั้นที่ 1 โดยวิเคราะห์จากข้อมูลที่เลือมา ดังนี้

1.2.1 age = อายุ

1.2.2 sex = เพศ

1.2.3 trestbps = อาการเจ็บหน้าอก

1.2.4 chol = ปริมาณคอเรสเตอรอน

1.2.5 fbs = น้ำตาลในเลือด

1.2.6 restecg = การวางตัวคลื่นไฟฟ้าให้ได้ผลสำเร็จ

1.2.7 thalach = อัตราการเต้นของหัวใจสูงสุดด้วยไฟฟ้า

1.2.8 exang = การออกกำลังกาย

1.2.9 slope = ความลาดชันของการออกกำลังกาย

1.2.10 thal = ข้อบกพร่อง

1.2.11 targt = เป็นโรคหัวใจ และ ไม่เป็นโรคหัวใจ

1.3 การจำแนกประเภทข่อมูล เปนขั้นตอนการทดสอบการจำแนกประเภทข่อมูล หลังจากทำการเลือกคุณลักษณะของข่อมูลที่มีความเหมาะสมและสอดคลอ่งมากที่สุด โดยทำการเปรียบเทียบคาเฉลี่ยความถูกตองการจำแนกประเภทข่อมูลของแตละกลุมการทดสอบ ดวยวิธีหาสมาชิกที่ใกลที่สุด (K-Nearest Neighbor) เบยอยางงาย และ Neural Network

4.2 เครื่องมือการวิจัย

2.1 RapidMiner Studio

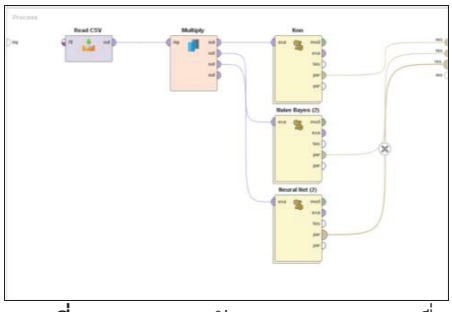
2.2 Excel

4.3 ประชากรและกลุ่มตัวอย่าง

3.1 กลุ่มตัวอย่าง คือ ข้อมูลที่นำมา 303 คน ข้อมูลนี้นำมาจาก (https://www.kaggle.com/ronitf/heart-disease-uci) Heart Disease UCI https://archive.ics.uci.edu/ml/datasets/Heart+Disease

4.4 การทดลองใช้โปรแกรม RapidMiner Studio

ผู้วิจัยดำเนินการทดลองใช้ โปรแกรม RapidMiner Studio ที่พัฒนาขึ้นเพื่อจำแนกข้อมูลโรคหัวใจด้วยเทคนิคเหมืองข้อมูล จำนวน 303 คน โดยมีเทคนิค 3 แบบ คือ K-Nearest Neighbor, Naïve Bayesian, Neural Network



ภาพที่ 1 ภาพแสดงตัวแบบประมวลผลเพื่อจำแนกข้อมูล RapidMiner Studio

accuracy: 82.17% +/- 7.51% (micro average: 82.18%)

	true 1	true 0	class precision
pred. 1	140	29	82.84%
pred. 0	25	109	81.34%
class recall	84.85%	78.99%	

ภาพที่ 2 ภาพแสดงตัวแบบประมวลผลเพื่อจำแนกข้อมูล Neural Network

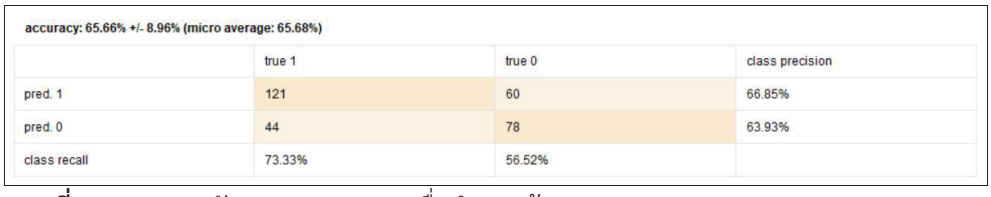
จากภาพที่ 2 จะเห็นว่า มีค่าความถูกต้อง 82.17%+/-7.51%

accuracy: 81.88% +/- 5.48% (micro average: 81.85%)

	true 1	true 0	class precision
pred. 1	139	29	82.74%
pred. 0	26	109	80.74%
class recall	84.24%	78.99%	

ภาพที่ 3 ภาพแสดงตัวแบบประมวลผลเพื่อจำแนกข้อมูล Naive Bayes

จากภาพที่ 3 จะเห็นว่า มีค่าความถูกต้อง 81.88%+/-5.48%



accuracy: 65.66% +/- 8.96% (micro average: 65.68%)

	true 1	true 0	class precision
pred. 1	121	60	66.85%
pred. 0	44	78	63.93%
class recall	73.33%	56.52%	

ภาพที่ 4 ภาพแสดงตัวแบบประมวลผลเพื่อจำแนกข้อมูล Knn

จากภาพที่ 4 จะเห็นว่า มีค่าความถูกต้อง 65.66%+/-8.69%

5. ผลการวิจัย

5.1 ผลการเปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทข้อมูล

ผู้วิจัยได้เปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนประเภทข้อมูล ผลการเปรียบเทียบการคัดเลือกคุณลักษณะในการคัดเลือกคุณลักษณะด วยเทคนิค 3 แบบ คือ K-Nearest Neighbor, Naïve Bayesian, Neural Network ดังนี้

ตารางที่ 1.1 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลโดย เทคนิค 3 แบบ คือ K-Nearest Neighbor, Naïve Bayesian, Neural Network

	K-Nearest Neighbor	Naïve Bayesian	Neural Network
ความแม่นยำ	65.66% +/- 8.96%	81.88% +/- 5.48%	**82.17% +/- 7.51%**

จากตารางที่ 1.1 การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล ประกอบด้วย K-Nearest Neighbor, Naïve Bayesian, Neural Network จะเห็นได้ว่า Neural Network มีความแม่นยำมากที่สุดคือ 82.17%/-7.51% รองลงมาคือ Naïve Bayesian 81.88%+/-5.48% และ K-Nearest Neighbor 65.66%+/-8.96%

6.อภิปรายผลการวิจัย

ผลการวิจัยครั้งนี้โดยพัฒนาและเปรียบเทียบตัวแบบการจำแนกทั้ง 3 เทคนิค ได้แก่ Naïve Bayes, k-NN, Neural Network ซึ่งผลการประเมินประสิทธิภาพตัวแบบ คือ Neural Network ซึ่งได้ค่าที่สูงที่สุดจากข้อมูลค่าความถูกต้อง (Accuracy) ได้ 82.17% จึงสรุปได้ว่า Neural Network เป็นตัวแบบที่เหมาะสมที่สุดที่จะนำมาใช้จำแนกข้อมูล

7. ข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยที่ใช้โปรแกรม RapidMiner Studio เพื่อนำมาประยุกต์ใช้ในการจำแนกข้อมูล (Performance Classification) โดยที่มีเครื่องมือต่าง ๆ ให้เลือกใช้ อาทิเช่น Naïve Bayes, k-NN,

Neural Network, Decision Tree โดยสามารถนำเครื่องมือเหล่านี้มาเปรียบเทียบการจำแนกเพื่อหาค่าความถูกต้องมากที่ฤด

เอกสารอ้างอิง

ฝนทิพย์ คูนแกว. (2555).**การสังเคราะห์โมเดลเพื่อการจำแนกตามข้อกำหนดของผู้ใช้**. มหาวิทยาลัยสุรนารี, นครราชสีมา

ไพศาล จันทรเจริญ , สุพจนเฮงพระพรหม และ ไกรุง เฮงพระพรหม. (2559).**การเปรียบเทียบประสิทธิภาพการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทขอมูลไมโครอารเรย**. สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยศาสตร์และเทคโนโลยีสารสนเทศมหาวิทยาลัยราชภัฏ นครปฐม

ธีรพงษ์ สังข์ศรี, (2557).**การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษา ใหม่โดยใชเทคนิคการทำเหมืองข้อมูล,** การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศครั้งที่10, กรุงเทพฯ

พยุง มีสัจ. (2555). **ระบบฟัซซีและโครงข่ายประสาทเทียม.** คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือกรุงเทพฯ

ซากี นิเซ็ง. (2553). **การกำหนดลักษณะและจำแนกมังคุดโดยโครงข่ายใยประสาทเทียม**. วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต. มหาวิทยาลัยวลัยลักษณ์. นครศรีธรรมราช.